



5

GRAPH DATA SCIENCE BASICS

EVERYONE
SHOULD KNOW

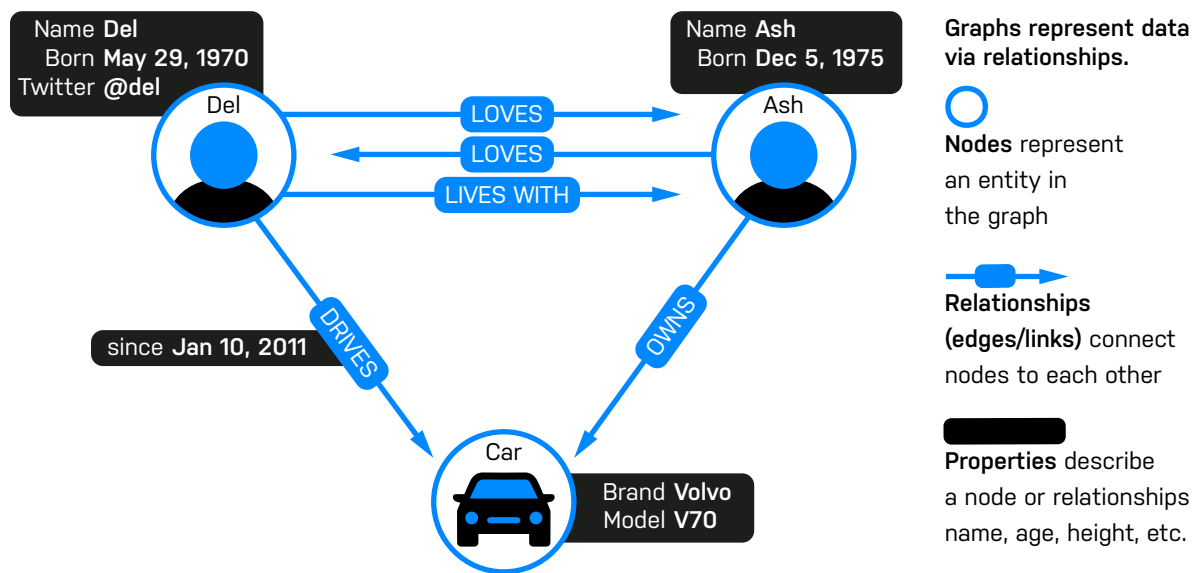
Why Graph Data Science and Why Now?

Commercial applications of graph data science are new and most data experts are still coming up to speed on how to best use it in their organizations. While some data experts studied graph theory, others have not had much exposure to it. Graph data science brings together graph analytics, statistics, and AI and ML techniques to improve their predictive and prescriptive models. This paper walks you through the basics of graph data science so you will feel confident knowing when to use it in your daily work.

1. What is a graph?

Before you can understand graph data science, you need to understand graphs. At its most fundamental, a graph is simply a different way of structuring data. Instead of rows and columns, like in a traditional, relational database table or dataframe, graphs use nodes (nouns) and relationships (verbs) as their primary structure.

In a graph, nouns – people, places, things, organizations – are nodes. The relationships between them are verbs: friends, works for, likes, and so on.



2. How is graph data science different from traditional analytics?

When you analyze data in tabular form, like in a relational database, you try to make sense of data points without a coherent way to analyze their connections. Ben Squire, a data scientist, compared traditional methods to "trying to solve a Rubik's Cube by only looking at one side." A graph gives data experts the ability to look at, understand, and analyze the connections between each data point. This gives data context that is impossible to get from a tabular data model. You can see how strong connections are, where groups of connections form, how important each connection is, and how connections influence one another.

Graph data science brings together graph statistics, analytics, and ML to put data in context and answer pressing questions.

Graph statistics, queries, and visualization drive exploration and insights. Graph statistics provide basic measures about a graph, such as the number of nodes and the distribution of relationships. Graph queries answer any question, no matter how deep, whether 6 or 600 degrees of separation. Graph visualization empowers data experts to see their data and explore patterns that bear further investigation.

Graph analytics builds on graph statistics by answering specific questions and gaining insights from connections in existing or historical data. Graph queries and algorithms are typically applied together in "recipes" during graph analytics, and the results are used directly for analysis.

Graph-enhanced ML is the application of graph data and analytics results to train ML models or support probabilistic decisions within an AI system. Graph statistics and analytics are often used in conjunction to answer certain types of questions about complex systems and the subsequent insights, applied to improve ML.

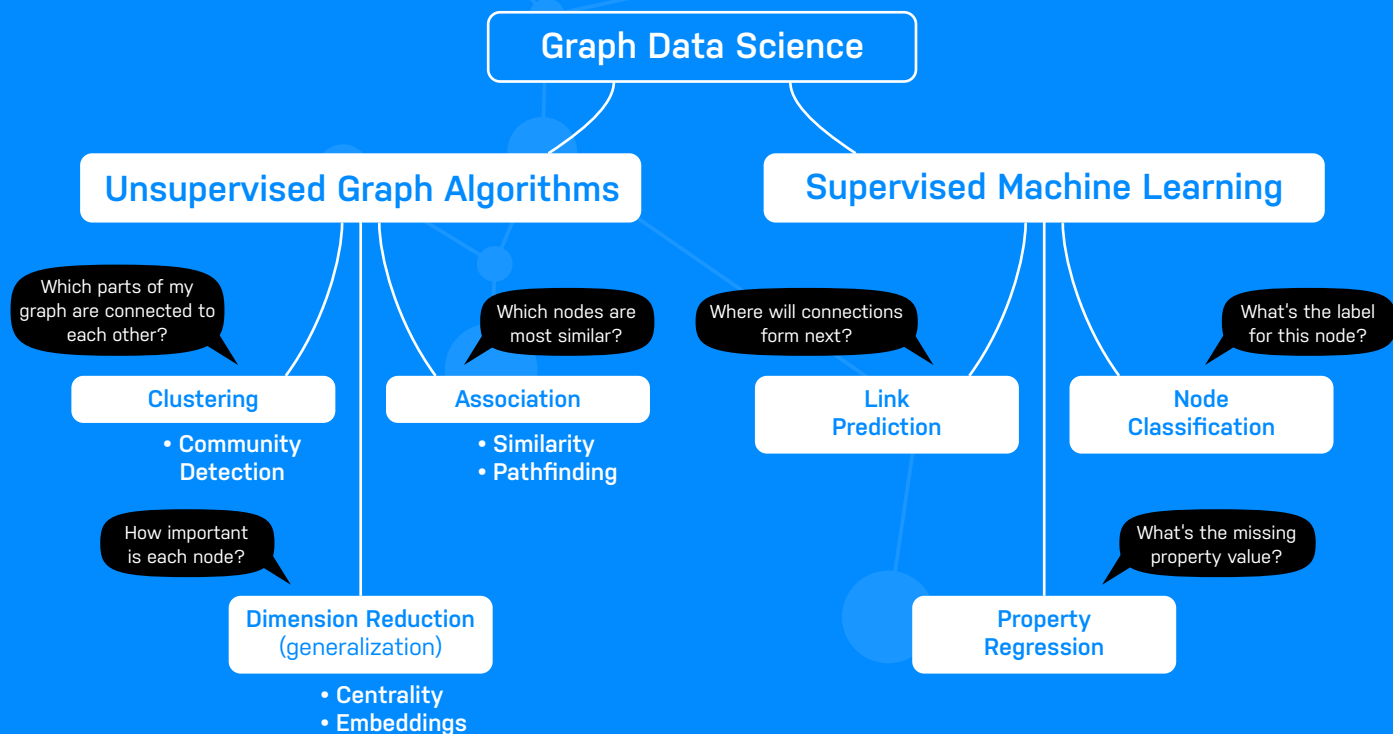
Quick Look: In-Graph ML Techniques That Do the Work for You

Interest in graph-based machine learning is exploding. By 2025, [Gartner](#) projects that graph technology will fuel 80% of data and analytics innovations, up from 10% in 2021. Here are a few advanced graph data science techniques your team may want to explore.

Link prediction fills in the blanks in your data and predicts changes in your graph's structure. Link prediction is a common machine learning task applied to graphs: training a model to learn where relationships should exist between pairs of nodes in a graph. You can think of link prediction as building a model to predict missing relationships in your dataset or relationships that are likely to form in the future. With graph data science, you can train supervised ML models based on the relationships and node properties in your graph to predict the existence – and probability – of relationships.

Node embedding transforms the topology and features of your graph into a low-dimensional vector representation of each node. These vectors, also called embeddings, can be used for exploratory data analysis, similarity measurements, and ML. Node embeddings can aggregate information about a node's position in the graph, its local neighbors, its centrality and influence, and in some cases, other numeric node properties.

Node classification models predict the class of nodes in your graph. A class can be a binary indicator, like whether a user account is engaged in fraud, or a multivalued indicator, such as which market segment a customer belongs to. Node classification models can be trained to predict which class nodes (including any new nodes) belong to. Node classification can incorporate a broad range of input features, including the network structure of your graph and properties from your source data.



3. Why do graphs matter?

Across an organization, every department can benefit from graph data science to answer questions like who or what is important, what should I do next, and what's unusual about this?

While use cases for graph data science span industries and lines of business – from life sciences to manufacturing – a few use cases are rapidly becoming the most popular among data scientists.



Anomaly and fraud detection

Anomaly detection across corporate networks can help to identify cybersecurity attacks and prevent data loss. The same strategy used to identify threat actors in a cybersecurity context can be used to detect fraud in banking, insurance, and government programs by analyzing the relationships and behaviors in your graph.

Those using graph data science to curb fraud have seen detection improvements of over 300%, saving millions every year.

Customer 360

Across the globe, businesses try to better understand their customers and improve lifetime value (LTV). With graph data science, customer knowledge can become more accurate and complete through entity resolution. This process looks at all the database entries and identifies duplicates. Creating a complete, master database entry for each customer instead of having multiple, incomplete entries improves LTV and deepens customer knowledge, allowing for optimized marketing programs and offers.

Those using graph data science for customer 360 can increase customer knowledge by 30%.

Recommendation engines

Recommendation engines became well known through Netflix and online shopping experiences. However, recommendation engines have uses across the business. From product development to human resources for retaining employees through upskilling training, recommendation engines power some of the most important parts of a business.

Supply chain management

Improving a supply chain leads to savings, not just in dollars, but also in carbon emissions. Every optimized route, perfect timing, and perfect delivery mean happier customers and less waste across time, infrastructure, and emissions. Graph data science helps optimize supply chain routes by finding the best path, balancing cost and efficiency with customer satisfaction and sustainability.

An ROI of tens of millions for one organization using graph data science for route optimization is nothing compared to the 60,000 tons of carbon emissions eliminated by using those optimized routes.

4. What are the big questions graph data science helps answer?

Graph data science helps you answer big questions to make critical business decisions. If you hear questions like these, you may benefit from using graph data science to answer them:

What's unusual?

To understand anomalies and hidden patterns in your graph, consider using **community detection**. Communities are clusters within your graph, and [community detection algorithms](#) can be used to discover and identify these clusters. Detecting communities helps you uncover unusual patterns, predict similar behavior, find duplicate entities, or simply prepare data for other analyses.

What should I recommend?

To build a recommendation engine, consider using similarity. **Similarity** identifies repeating patterns in your graph. [Similarity algorithms](#) employ set and distance comparisons to score how alike individual nodes are based on their neighbors or properties. This approach is used in applications such as personalized recommendations and developing categorical hierarchies.

What content is the most important? Who is the most influential? Where is our biggest risk?

To understand what's important, use **centrality** algorithms. Centrality metrics like [PageRank](#) help you identify what's important. [Centrality algorithms](#) reveal which nodes are important based on graph topology. They identify influential nodes based on their position in the larger network, including their connections. These algorithms are used to infer group dynamics such as credibility, cascading vulnerability, and bridges between groups.

What is the optimal route?

To understand route optimization, use **pathfinding**. [Pathfinding algorithms](#) find the best routes across your connected data. Pathfinding algorithms are foundational to graph analytics and find the most efficient or shortest paths to traverse between nodes. They can be used to understand complex dependencies and evaluate routes for uses such as physical logistics and least-cost call or IP routing.

5. Who uses graph data science tools?

Anyone who works with data can take advantage of graph data science to help find answers to difficult questions. Here are some of the roles currently using graph data science.



Data scientists are typically the primary user of graph data science tools because they are practitioners of data science with deep knowledge of algorithms and models.



Machine learning engineers work to scale, improve, embed, integrate or operate machine learning models that are developed by data scientists.



Data engineers identify trends in data sets, create data pipelines, and leverage graph algorithms to transform and enrich graph data to make it more useful to the organization and other data experts.



Citizen data scientists apply graph data science techniques without the need for deep analytics and ML expertise, using a low-code/no-code experience.



Business data analysts visualize, build upon, and report on the analysis done by a data scientist for business users.



Application developers and software architects who are learning about graph databases often use graph queries, typically using [Cypher query language](#). They learn graph algorithms as they seek to analyze patterns across their graph and use those techniques in their applications. (Graph data science is often the secret sauce in differentiating applications.)

Stay Ahead of the Curve

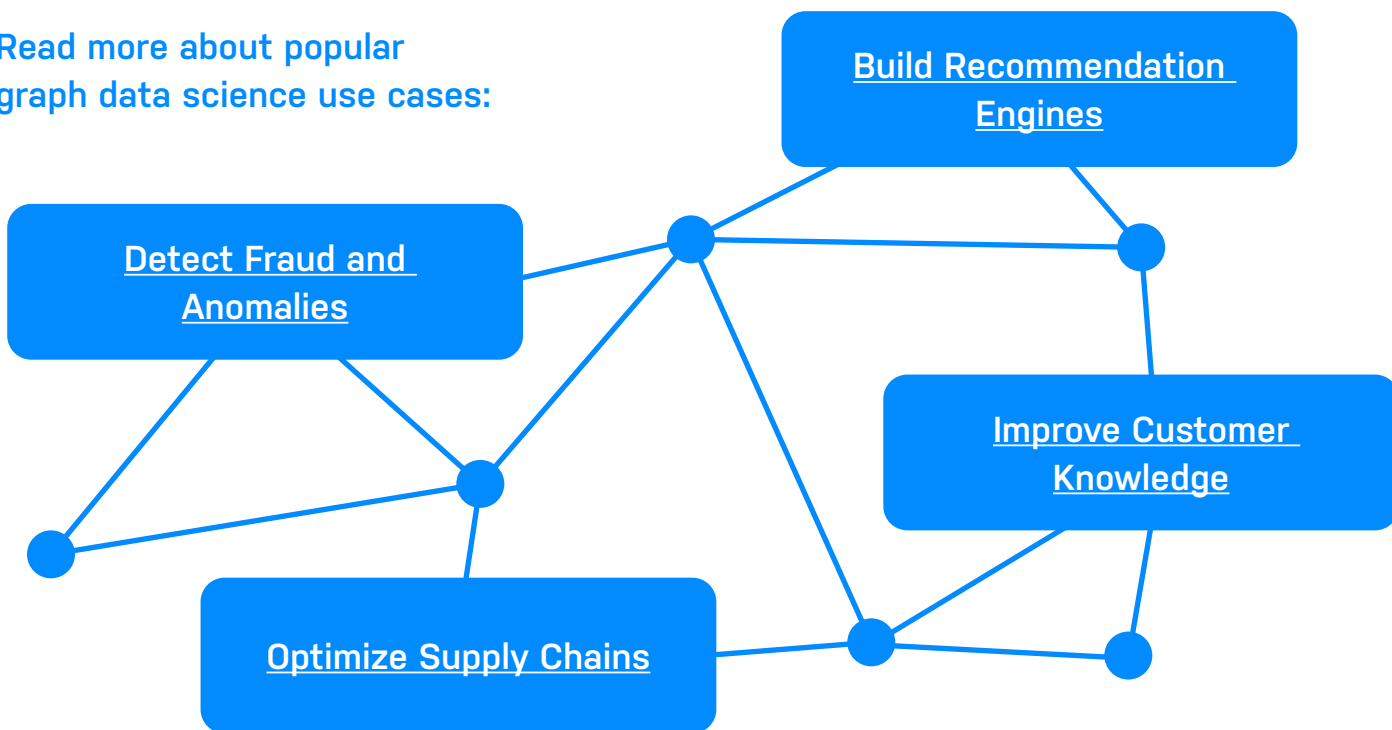
Commercial applications of graph data science are new and data experts are still coming up to speed on how to best use graph data science in their organizations. When considering use cases, data experts and data scientists should keep in mind that graph data science helps answer big questions like what's important, what's unusual, and what's next. Using this framing, it is easier to identify opportunities to use graph data science to improve models and make predictions.

Organizations of all sizes, all industries, and within all departments are using graph data science to make recommendations, identify anomalies and find fraudsters, improve customer knowledge, and optimize supply chains.

So what will you do with graph data science?

Ready to try it out? Activate [a free graph data science sandbox from Neo4j](#) with prepopulated data for common scenarios.

Read more about popular graph data science use cases:



Neo4j is the world's leading graph data platform. We help organizations – including [Comcast](#), [ICIJ](#), [NASA](#), [UBS](#), and [Volvo Cars](#) – capture the rich context of the real world that exists in their data to solve challenges of any size and scale. Our customers transform their industries by curbing financial fraud and cybercrime, optimizing global networks, accelerating breakthrough research, and providing better recommendations. Neo4j delivers real-time transaction processing, advanced AI/ML, intuitive data visualization, and more. Find us at [neo4j.com](#) and follow us at [@Neo4j](#).

Questions about Neo4j? Contact us around the globe:

info@neo4j.com
neo4j.com/contact-us